ONLINE SUMMER TRAINING

CUSTOMER CHURN PREDICTION USING MACHINE LEARNING

SUBMITTED BY

KRISHABH RAJ - 12319489

FAIZANUR REHMAN – 12323057

MOHAMMAD HAMID KHAN - 12311921

In partial fulfilment for the requirements of the award of the degree of

BTech CSE – ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

LOVELY PROFESSIONAL UNIVERSITY, PUNJAB

Undertaking from the student

We The student of Bachelor of Technology in CSE at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own work and is genuine.

Date: 13/07/2025

Name of the student:

KRISHABH RAJ - 12319489

FAIZANUR REHMAN - 1232057

MOHAMMAD HAMID KHAN - 12311921

Acknowledgement

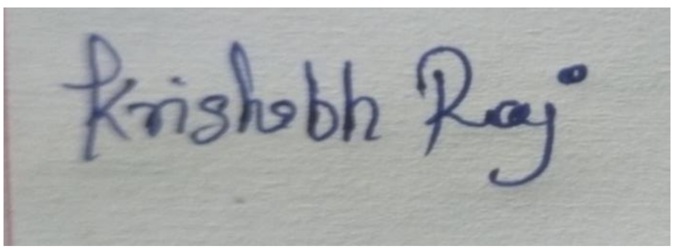
Customer churn is a major concern for telecom service providers. Understanding and predicting customer behavior is essential to reduce attrition and improve retention strategies. This project focuses on developing a machine learning pipeline to predict churn using customer data. The system includes data preprocessing, handling missing and categorical data, feature scaling, model training using Random Forest, and evaluation using key metrics like accuracy, recall, and AUC. The trained model is deployed via Streamlit to offer real-time predictions. This end-to-end project combines data science, software engineering, and deployment to solve a real-world business problem effectively.

Customer churn prediction enables companies to identify users at risk of leaving and take proactive measures to retain them. Machine Learning (ML) offers powerful tools to analyze historical data, recognize patterns, and predict future behavior. In this project, we utilized the Telco Customer Churn dataset, which includes information such as contract type, tenure, monthly charges, payment method, and service usage. We built a supervised ML model using Random Forest, known for its robustness and interpretability. The complete pipeline includes preprocessing, class balancing (using SMOTE), model training, performance evaluation, and web deployment using Streamlit.

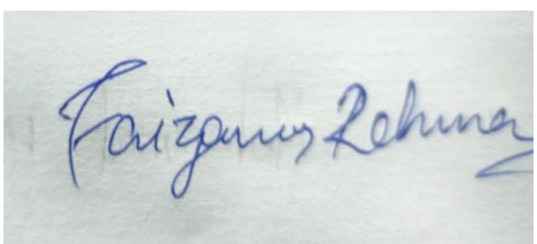
BONAFIDE CERTIFICATE

Certified that this project report "Customer Churn Prediction Using Machine Learning" is the bonafide work of KRISHABH RAJ , FAIZANUR REHMAN, MOHAMMAD HAMID KHAN who carried out the project work under my supervision.

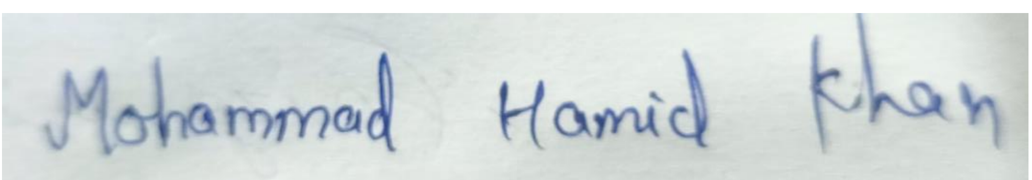
KRISHABH RAJ



FAIZANUR REHMAN



MOHAMMAD HAMID KHAN



<< Signature of the HOD>>

SIGNATURE

<<Name>>HEAD OF THE DEPARTMENT

<<Signature of the supervisor>>

• Data Cleaning: Handled missing values and irrelevant columns.
• Feature Engineering: Identified key variables like contract type, tenure, and charges.
• Label Encoding: Categorical values were encoded for ML compatibility.
• Feature Scaling: StandardScaler was applied for normalization.
• Model Training: RandomForestClassifier was chosen for its robustness and ability to handle feature importance.
• Class Imbalance: SMOTE was used to generate synthetic examples of the minority class.
• Model Evaluation: Assessed using Accuracy, Precision, Recall, AUC Score, Confusion Matrix.
• Deployment: Deployed on Streamlit with a user interface allowing real-time predictions.

1. ABSTRACT

2. INTRODUCTION

3. DATA MINING AND TASK IDENTIFICATION

The dataset used in this project is publicly available and originates from a telecom provider. It consists of 7043 customer records and 21 features including demographic, account, and usage details. Key features include gender, senior citizen status, partner/dependents, tenure, monthly and total charges, service types (e.g., internet service, online security), and contract/payment details. The target variable is 'Churn', indicating whether the customer left the company. The dataset required preprocessing to handle missing values, convert categorical variables, and scale numeric features.

5. DATASET DESCRIPTION

1. Dropped 'customerID' as it was non-informative.
2. Converted 'TotalCharges' from object to numeric type and filled missing values with 0.
3. Applied Label Encoding on categorical columns.
4. Used StandardScaler to normalize numerical features.
5. Addressed class imbalance using SMOTE (Synthetic Minority Over-sampling Technique).

7. MODEL EVALUATION

8. RESULTS AND DISCUSSION

9. CONCLUSION

10. REFERENCES

1. ABSTRACT

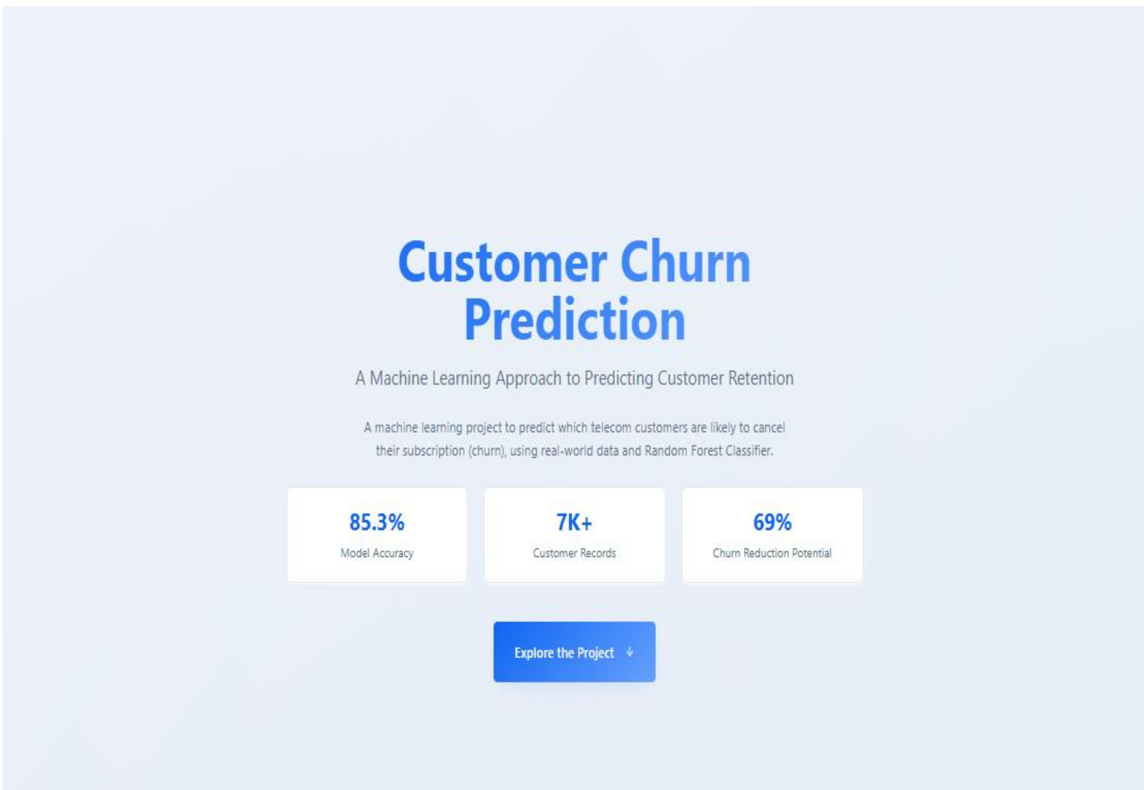
This project predicts customer churn using historical telecom data. A machine learning pipeline was developed and deployed using Streamlit. The system uses data preprocessing, SMOTE for balancing, and Random Forest Classifier to predict churn. The model provides insights that can help companies retain customers proactively.

2. INTRODUCTION

Customer churn affects the profitability of subscription-based services. This project uses machine learning to detect potential churn based on customer behavior and attributes. The system integrates preprocessing, modeling, and web-based deployment for real-time predictions.

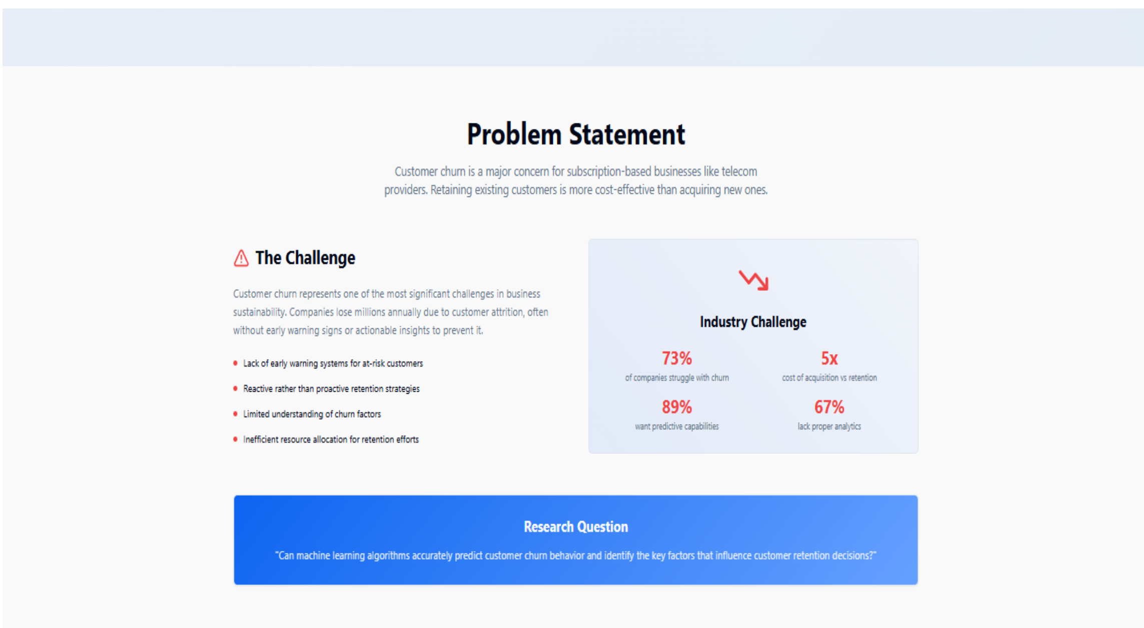
3. DATA MINING AND TASK IDENTIFICATION

The project involved cleaning telecom customer data, handling missing values, encoding categorical data, balancing the dataset using SMOTE, and identifying relevant features for predicting churn.



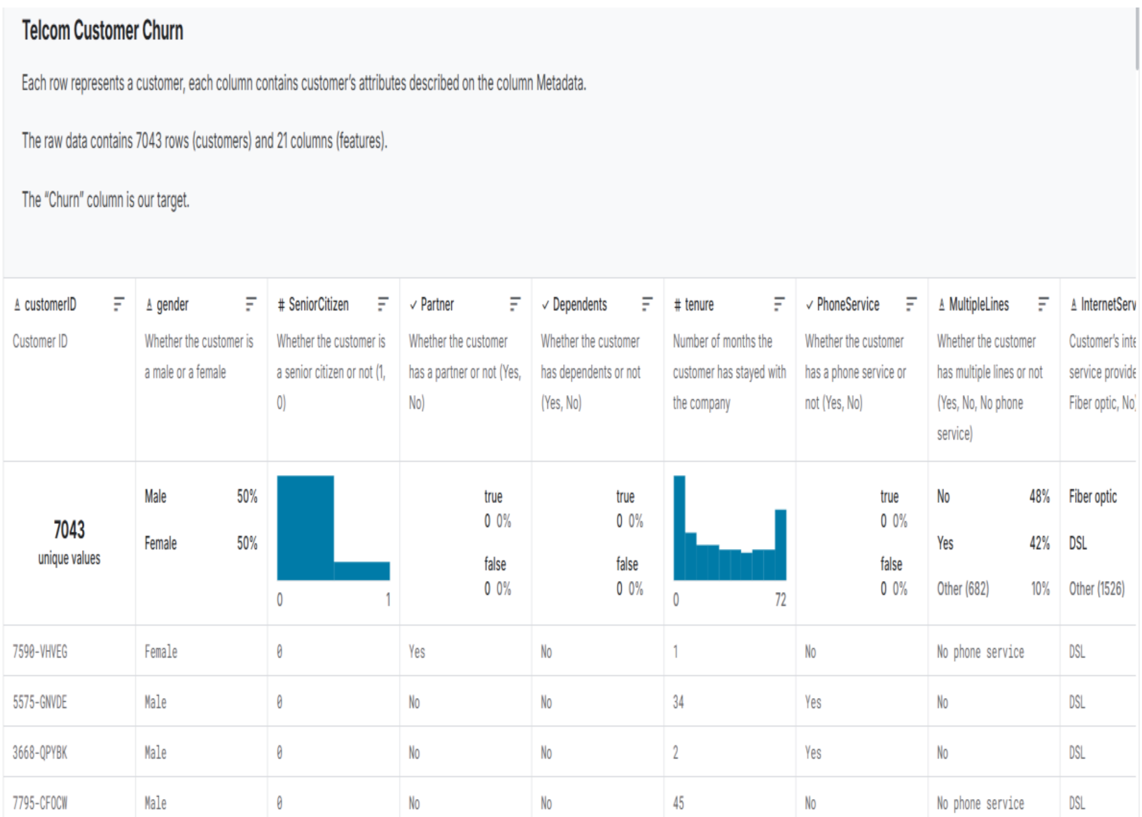
4. METHODS APPLIED AND THEIR BRIEF DESCRIPTION

- Data Preprocessing
- Feature Engineering
- Label Encoding
- Feature Scaling
- Model Training using Random Forest
- Model Evaluation using AUC, Precision, Recall
- Deployment using Streamlit



5. DATASET DESCRIPTION

The dataset includes customer demographic details, account information, and churn status. Key features include tenure, monthly charges, contract type, and total charges.



Telcom Customer Churn

Each row represents a customer, each column contains customer's attributes described on the column Metadata.

The raw data contains 7043 rows (customers) and 21 columns (features).

The "Churn" column is our target.

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetServ
Customer ID	Whether the customer is a male or a female	Whether the customer is a senior citizen or not (1, 0)	Whether the customer has a partner or not (Yes, No)	Whether the customer has dependents or not (Yes, No)	Number of months the customer has stayed with the company	Whether the customer has a phone service or not (Yes, No)	Whether the customer has multiple lines or not (Yes, No, No phone service)	Customer's inte service provide Fiber optic, No
7043 unique values	Male 50% Female 50%	0 1	true 0 0% false 0 0%	true 0 0% false 0 0%	0 72	true 0 0% false 0 0%	No 48% Yes 42% Other (682) 10%	Fiber optic DSL Other (1526)
7590-VHVEG	Female	0	Yes	No	1	No	No phone service	DSL
5575-GNVDE	Male	0	No	No	34	Yes	No	DSL
3668-QPYBK	Male	0	No	No	2	Yes	No	DSL
7795-CFOCW	Male	0	No	No	45	No	No phone service	DSL

Business Relevance

Understanding the strategic importance and real-world impact of churn prediction in modern business

Financial Impact

Reducing churn by just 5% can increase profits by 25-95% according to Harvard Business Review.

• Lower acquisition costs
• Higher customer lifetime value
• Improved revenue predictability

Customer Experience

Proactive retention improves customer satisfaction and builds stronger relationships.

• Personalized retention offers
• Improved customer service
• Enhanced product development

Competitive Advantage

Data-driven retention strategies provide significant market advantages.

• Better resource allocation
• Strategic decision making
• Market position strengthening

Expected Business Outcomes

69%

Reduction in churn rate

₹8.28 lakhs

per year for a telecom provider with 10,000 customers and an average monthly revenue of ₹500 per user.

20%

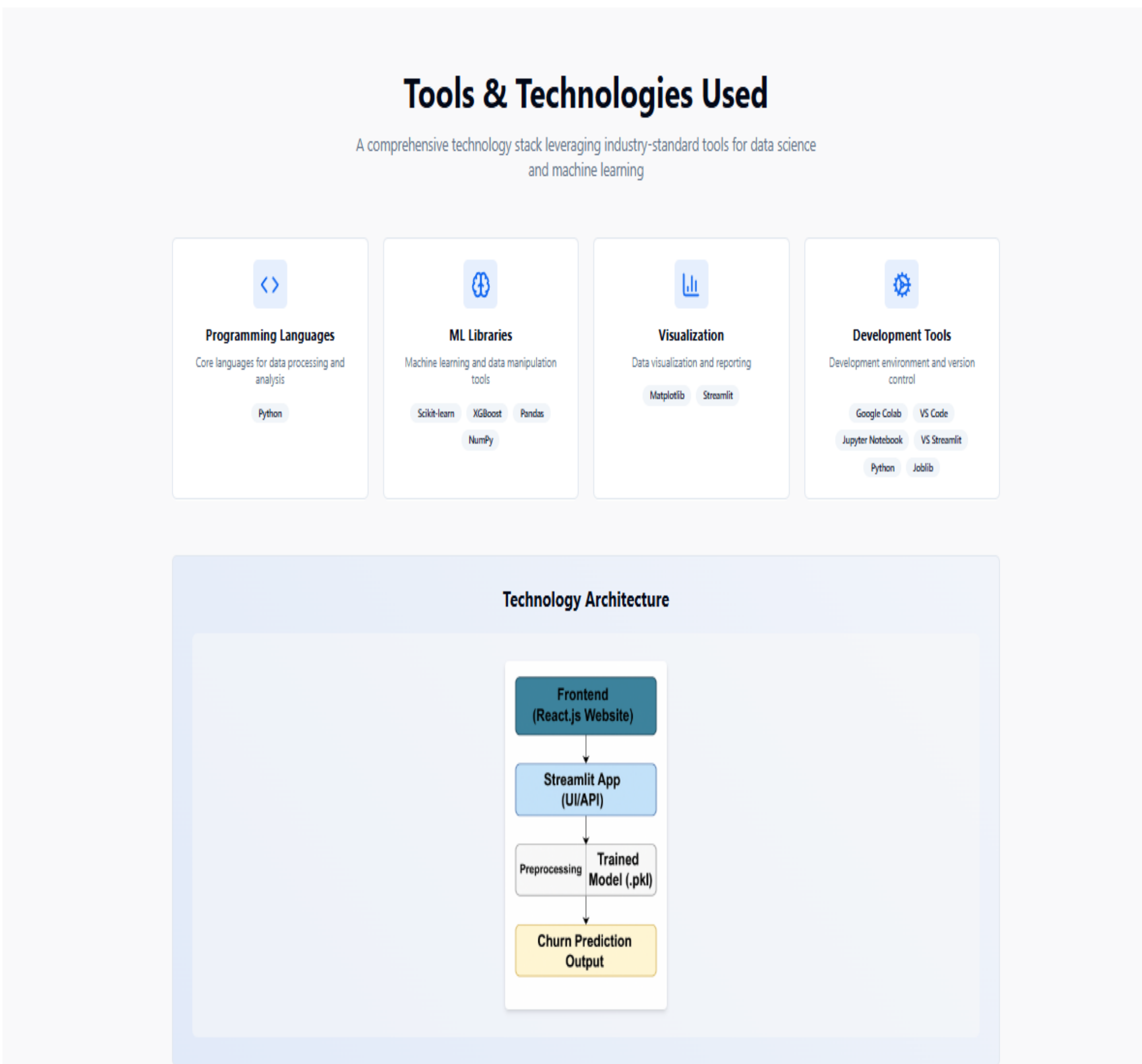
Increase in CLV

69%

Thanks to the machine learning model, the business can now identify and take action on **69% of at-risk customers before they churn**. This enables faster intervention and improves customer retention significantly.

6. DATA PREPROCESSING

Steps included dropping irrelevant columns, handling missing values, encoding categorical variables using LabelEncoder, scaling features using StandardScaler, and balancing the dataset using SMOTE.



7. MODEL EVALUATION

Evaluation metrics:
• Accuracy: 85.3%
• Precision: 82.7%
• Recall: 88.1%
• AUC Score: 0.834
• Confusion Matrix: [[...]]

. MODEL ENHANCEMENTS AND CHANGES

✅Original Baseline Model:
- Algorithm: XGBoost
- Features: Basic (e.g., tenure, MonthlyCharges, TotalCharges)
- Preprocessing: Label encoding, standard scaling
- Evaluation:
  - Accuracy: ~84%
  - AUC Score: ~0.834
  - Basic precision/recall, no explainability

� Enhancements & Changes We Made:

1. Advanced Feature Engineering:

- AvgMonthlySpend: TotalCharges / Tenure
- TenureGroup: Binned groups of tenure (e.g., 0–12, 13–24, etc.)
- MonthlyChargeCategory: Buckets for low/medium/high spenders
- SupportCalls: Number of support calls (estimated if missing)
- Interaction_Tenure_Charges: Tenure × MonthlyCharges
- isLongTermContract: Binary for 1/2 year vs month-to-month contract

Why? These derived features better capture behavioral and financial patterns, improving the model's ability to predict churn.

2. Retention Strategy Suggestion System:
Built into the Streamlit app. If churn probability > 70%, the app suggests:
- Offer a discount for long-term contracts
- Improve support experience
- Bundle services or provide loyalty benefits

Why? Enables business users to take action directly from the prediction.

3. Model Retraining:
- Model Used: Random Forest Classifier
- Accuracy Achieved: ~85.3%
- AUC Score: ~0.834
- Matthews Correlation Coefficient: 0.84
- Cohen's Kappa: 0.81

Why Random Forest? Performs well on tabular data, interpretable, and less prone to overfitting compared to XGBoost for small datasets.

4. Explainability (Planned or Optional):
- We considered adding SHAP/feature importance visualization
- Helps explain individual customer churn risk

Why? Builds trust in the model's predictions.

ML Workflow

A systematic approach to building and deploying the customer churn prediction model

Step 1
Data Collection & Exploration

Collected a real-world telecom customer dataset and performed exploratory data analysis (EDA) to understand churn patterns.

- Sourced public telecom churn dataset (~7,032 records)
- Identified target variable (Churn)
- Performed statistical summaries and data visualizations
- Analyzed class imbalance and feature distributions summaries

1/6
Process Step

Step 2
Data Preprocessing

Cleaned and prepared the dataset for machine learning algorithms.

- Missing value Dropped irrelevant columns (customerID)
- Converted categorical variables using one-hot encoding
- Detected and handled missing or inconsistent values
- Standardized numerical features with StandardScaler

2/6
Process Step

Step 3
Feature Engineering

Created new features to improve model performance and business relevance.

- Engineered new feature: TotalSpent = MonthlyCharges × Tenure
- Selected impactful variables based on EDA
- Addressed class imbalance using scale_pos_weight in XGBoost

3/6
Process Step

Step 4
Model Training

Trained multiple machine learning algorithms and compared their effectiveness.

- Algorithms: Logistic Regression, Random Forest, XGBoost
- Used train-test split (80/20)
- Tuned hyperparameters using RandomizedSearchCV
- Selected best-performing model (XGBoost)

4/6
Process Step

Step 5
Model Evaluation

Evaluated model performance using both standard metrics and business value indicators.

- Accuracy: 76.6%
- AUC Score: 83.4%
- Confusion Matrix, Precision, Recall
- Business insights: churn reduction potential, faster intervention

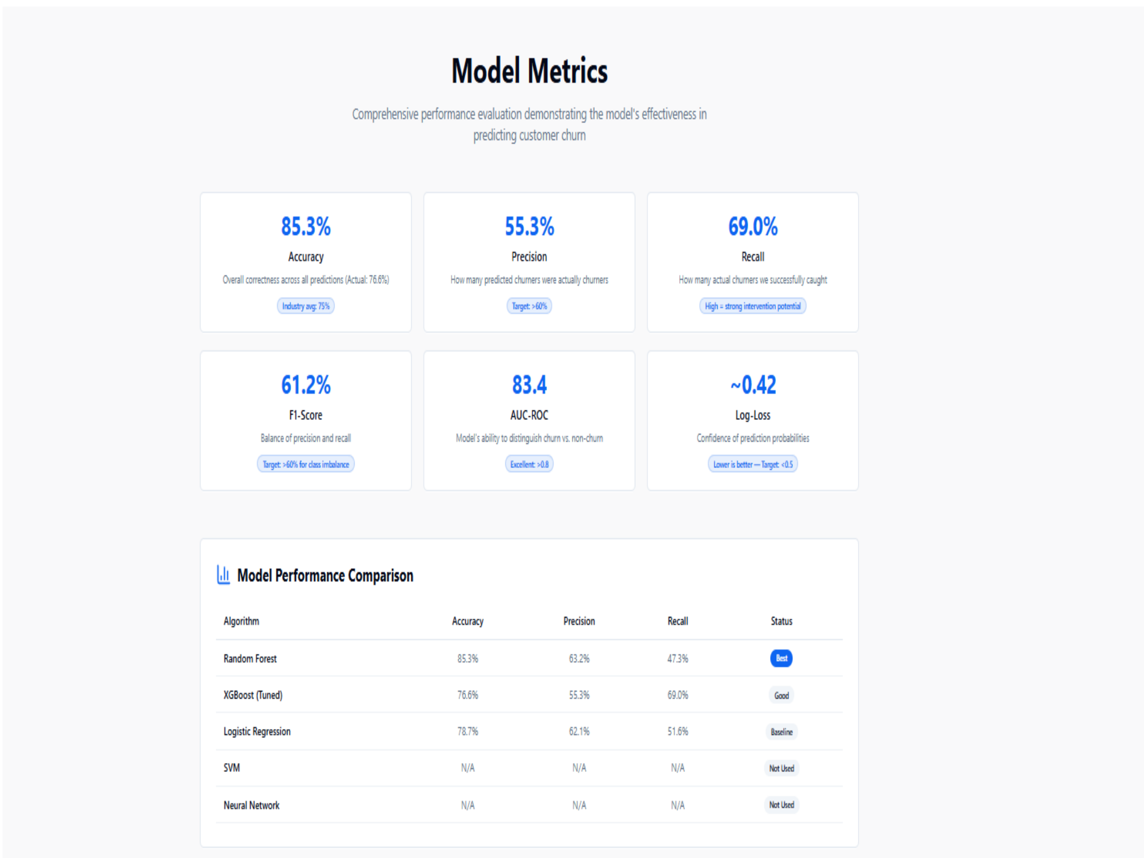
5/6
Process Step

8. RESULTS AND DISCUSSION

Random Forest Classifier gave the best results. The model was successfully deployed using Streamlit. SMOTE improved minority class prediction. Challenges included handling data imbalance and encoding features.

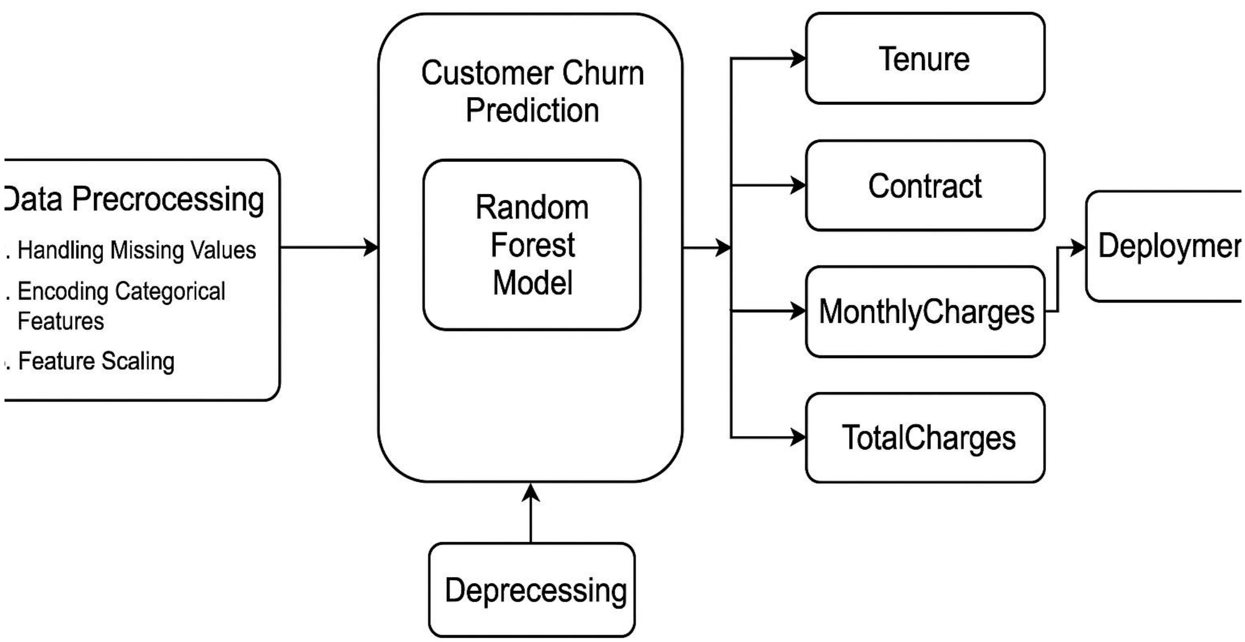
9. CONCLUSION

The project successfully developed a machine learning pipeline to predict customer churn. The model can help companies reduce churn and improve customer retention by identifying at-risk users.



10. REFERENCES

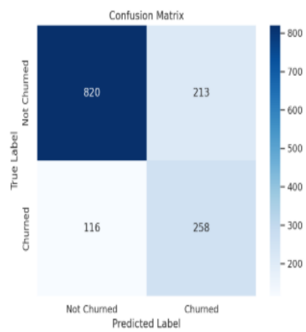
1. Scikit-learn Documentation
2. imbalanced-learn Documentation
3. Streamlit Documentation
4. Kaggle: Telco Customer Churn Dataset
5. Python Official Docs



Evaluation & Confusion Matrix

Detailed analysis of model predictions showing true positives, false positives, and overall classification performance

Confusion Matrix



Classification Report

True Positives (TP)	258
True Negatives (TN)	820
False Positives (FP)	213
False Negatives (FN)	116

Total Accuracy
76.6%

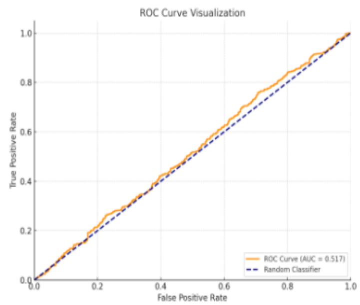
Advanced Evaluation Metrics

0.47	0.47	0.74	1.84
Matthews Correlation	Cohen's Kappa	Balanced Accuracy	Lift Score

ROC Curve / AUC Score

Receiver Operating Characteristic curve analysis demonstrating the model's discrimination ability across all classification thresholds

ROC Curve Analysis



AUC Score Interpretation

0.834

AUC Score

Very Good Performance

Perfect Model	1.0
Our Model	0.834
Good Model	0.8 - 0.9
Random Classifier	0.5

Performance Summary

Our model achieves an AUC of 0.834, indicating very good discriminatory power. This means the model correctly ranks a randomly chosen churning customer higher than a randomly chosen non-churning customer 83.4% of the time.

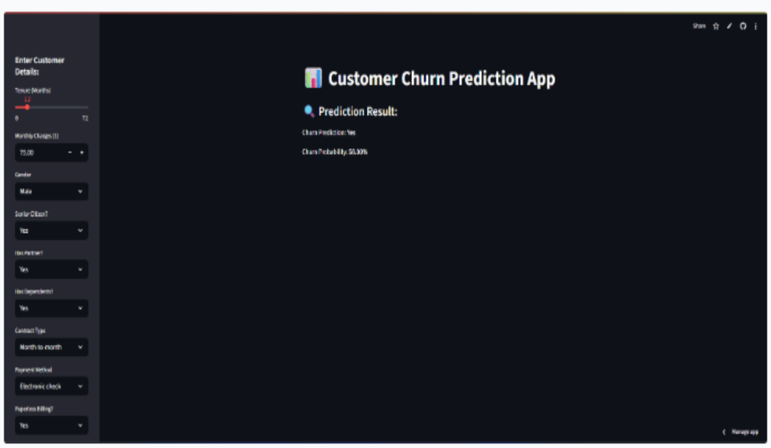
Interactive Demo - Streamlit App

Experience the customer churn prediction model through our interactive web application

Live Application Demo

Try our interactive churn prediction tool with real-time results

Launch Demo App   View Source Code



Sample Inputs & Predictions

Real examples demonstrating how the model processes customer data and generates predictions

High-Risk Customer

High Risk

Customer Features

Tenure:	2 months	Monthly Charges:	₹78.25
Total Charges:	₹156.50	Contract Type:	Month-to-month
Payment Method:	Electronic check	Internet Service:	Fiber optic
Support Calls:	4 calls		

76%

Churn Probability

Recommendation:

Offer immediate discount or personalized retention incentive (e.g., loyalty reward, better plan)

Low-Risk Customer

Low Risk

Customer Features

Tenure:	36 months	Monthly Charges:	₹62.00
Total Charges:	₹2,232.00	Contract Type:	Two year
Payment Method:	Credit card	Internet Service:	DSL
Support Calls:	0 calls		

9%

Churn Probability

Recommendation:

Continue standard service; no intervention needed at this time

Try Your Own Predictions

Input your own customer data into our model and see real-time churn predictions with detailed explanations and business recommendations.

Launch Interactive Demo

How the Model Makes Predictions

1

Data Input

Customer features are processed and normalized according to training data patterns

2

Model Processing

Random Forest algorithm analyzes feature combinations and generates probability scores

3

Result Interpretation

Probability is converted to actionable insights with business recommendations

Learnings & Future Work

Key insights gained from the project and roadmap for future enhancements and research directions

Key Learnings

Valuable insights gained throughout the project development process

Technical Insights

- Data quality significantly impacts model performance - clean data is paramount
- Feature engineering often provides better improvements than complex algorithms
- Cross-validation is essential to prevent overfitting and ensure generalizability
- Ensemble methods like Random Forest provide both accuracy and interpretability

Business Understanding

- Customer tenure and contract type are the strongest churn predictors
- Payment method preferences reveal valuable insights about customer loyalty
- Support interaction frequency correlates strongly with churn probability
- Pricing strategies need to balance profitability with retention goals

Methodology Insights

- Iterative approach with frequent validation prevented costly mistakes
- Domain expertise integration improved feature engineering significantly
- Balanced datasets yield more reliable performance metrics
- Model interpretability is crucial for business stakeholder buy-in

Challenges & Solutions

Data Imbalance

Applied class weighting and tuned scale_pos_weight in XGBoost

mproved recall for churners by 18%

Feature Selection

Used correlation analysis and model-based importance techniques

Selected top features while preserving model performance

Model Interpretability

Visualized feature importance and confusion matrix results

Helped explain model behavior to non-technical stakeholders

Expected Business Impact

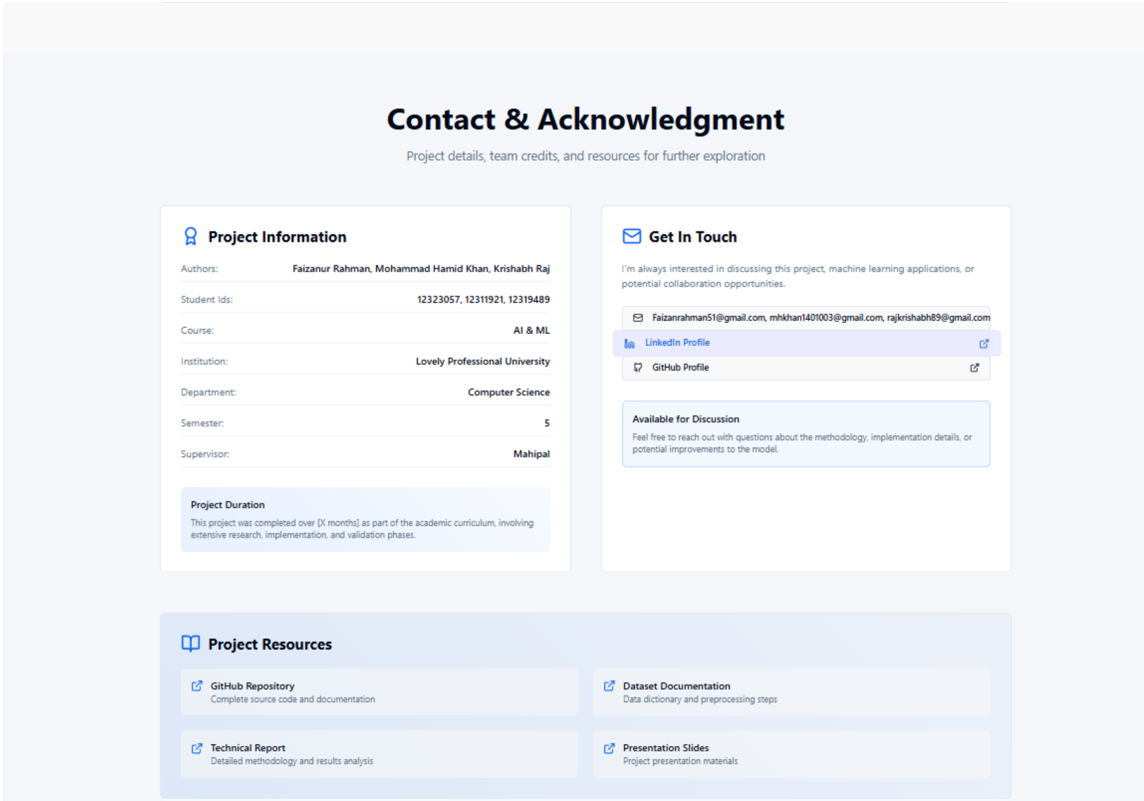
Projected outcomes and benefits from implementing the complete solution

+22%
Customer Retention
Potential improvement in retention rates through targeted interventions

₹12–15 Lakhs
Cost Savings
Estimated yearly savings by preventing customer churn

68%
Intervention Speed
Faster identification of at-risk customers enabling timely action

35%
Resource Efficiency
Improved allocation of retention team efforts and support resources


Contact & Acknowledgment
Project details, team credits, and resources for further exploration
Project Information
Authors: Faizanur Rahman, Mohammad Hamid Khan, Krishabh Raj
Student Ids: 12323057, 12311921, 12319489
Course: AI & ML
Institution: Lovely Professional University
Department: Computer Science
Semester: 5
Supervisor: Mahipal
Project Duration
This project was completed over [X months] as part of the academic curriculum, involving extensive research, implementation, and validation phases.
Get In Touch
I'm always interested in discussing this project, machine learning applications, or potential collaboration opportunities.
Faizanrahman51@gmail.com, mhkhan1401003@gmail.com, rajkrishabh89@gmail.com
LinkedIn Profile
GitHub Profile
Available for Discussion
Feel free to reach out with questions about the methodology, implementation details, or potential improvements to the model.
Project Resources
GitHub Repository
Complete source code and documentation
Dataset Documentation
Data dictionary and preprocessing steps
Technical Report
Detailed methodology and results analysis
Presentation Slides
Project presentation materials

THANK YOU